Environmental Modelling & Software 27-28 (2012) 52–61

Contents lists available at SciVerse ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

openair — An R package for air quality data analysis

David C. Carslaw [a,*], Karl Ropkins [b]

[a] King's College London, Environmental Research Group, Franklin Wilkins Building, 150 Stamford Street, London SE1 9NH, UK
[b] Institute for Transport Studies, University of Leeds, LS2 9JT, UK

ARTICLE INFO

Article history:
Received 9 February 2011
Received in revised form
12 September 2011
Accepted 26 September 2011
Available online 20 October 2011

Keywords:
Air pollution
Dispersion modelling
Reproducibility
Open source software
R

ABSTRACT

openair is an R package primarily developed for the analysis of air pollution measurement data but which is also of more general use in the atmospheric sciences. The package consists of many tools for importing and manipulating data, and undertaking a wide range of analyses to enhance understanding of air pollution data. In this paper we consider the development of the package with the purpose of showing how air pollution data can be analysed in more insightful ways. Examples are provided of importing data from UK air pollution networks, source identification and characterisation using bivariate polar plots, quantitative trend estimates and the use of functions for model evaluation purposes. We demonstrate how air pollution data can be analysed quickly and efficiently and in an interactive way, freeing time to consider the problem at hand. One of the central themes of *openair* is the use of conditioning plots and analyses, which greatly enhance inference possibilities. Finally, some consideration is given to future developments.

1. Introduction

1.1. Background

Worldwide an enormous and growing amount of air pollution data is collected. In Europe for example, the *Airbase* database consists of over 7000 monitoring sites, which still only represent a fraction of the total amount of data available (http://www.eea.europa.eu/themes/air/airbase). These data are collected for many reasons including to provide information for testing compliance against air quality standards, guidelines and regulations, and for research purposes. Most data are probably collected to monitor progress towards meeting legislative limits, such as those imposed by the European Union. Much, if not most of the data are analysed in superficial ways e.g. to confirm whether an annual mean concentration is above or below a stated threshold value. This situation represents a considerable missed opportunity. Experience shows that considerably more information can be gleaned from the analysis of data through the application of innovative data analysis techniques.

There are however several barriers that prevent the widespread use of more insightful analysis. First, a coherent set of data analysis tools for air pollution purposes does not exist. Second, many users are not aware of the tools available or how to apply them. Third, even where these tools are available, they are often only accessible through a variety of expensive, specialist proprietary software, resulting in an inconsistent and fragmented approach to data analysis. If these barriers can be overcome, there are many potential benefits including: a more comprehensive evidence base to support decision making, identification of the factors controlling pollutant concentrations (including the discovery of unexpected influences), an improvement in the quality of analyses undertaken and the enhanced validation of environmental models. There are also potential economic benefits through better decision making and targeted actions to manage air pollution.

This paper describes the development of a set of air pollution analysis tools using the R statistical software (R Development Core Team, 2011). First we describe the principal aims of the software and provide some background information on how its design was approached. Second, we provide a series of examples of usage with the aim of showing how inferences can be drawn from data. These examples include importing air pollution data, discovering and characterising source characteristics, quantifying trends and applying the tools to model output to understand model performance. Other recent examples of work that does make use of air quality data analysis techniques to derive more insight includes van Velzen and Segers (2010); Appel et al. (2011). In the case of Appel et al. (2011) a range of analysis techniques have been developed to better understand regional air quality model performance, which is an issue we also address in the current paper.

[*] Corresponding author.
E-mail address: david.carslaw@kcl.ac.uk (D.C. Carslaw).

1364-8152/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2011.09.008

2. Software design and characteristics

2.1. Development aims

The software development had several aims that were to a large extent governed by the barriers discussed in Section 1.1. The first decision was to choose a development framework that would simultaneously allow a high level of numerical and graphical capability, that was cross-platform and which the tools could be made readily available internationally. Second, in order to ensure maximum uptake and participation in the project, there was a strong desire to use open-source software that could be freely distributed. The use of open-source software was important from a transparency perspective and it was thought that users and policy-makers would favour an open system where the code and techniques were open to full scrutiny. The other important aspect of using open-source software is that it is more likely that the international air pollution community would participate in its development and contribute to extending its capabilities.

2.2. The R project

R (www.r-project.org) is an open-source programming environment which is gaining rapidly in usage across a many wide range of disciplines (R Development Core Team, 2011). It is an interpreted language that offers excellent interactive analysis capabilities and is ideal for the rapid development of statistical and data analysis applications. One of the principal advantages of R for the *openair* project is that it is free and open-source; thus overcoming some of the barriers noted in Section 1.1. It is very robust and works on a wide range of platforms including Microsoft Windows, Apple OS X and Linux.

For air pollution purposes, R represents the ideal system with which to work. Core features such as effective data manipulation, data/statistical analysis, high quality graphics and visualisation lend themselves to analysing air pollution data. While R began and is probably best known as a statistical programming language, its strong capability of working with data in general means that it has a much wider impact, covering a very wide range of disciplines. The ability to develop one's own analyses, invent new approaches etc. using R means that advanced tools can be quickly developed for specific purposes. The use of R ensures that analyses and graphics are not constrained to “off the shelf” tools. These tools will often contain functionalities that are either part of the R base system or that exist through specific packages.

Another key strength of R is its package system. The base software, which is in itself highly capable (e.g. offering for example linear and generalized linear models, non-linear regression models, time series analysis, classical parametric and non-parametric tests, clustering and smoothing), has been greatly extended by additional functionality. Packages are available to carry out a wide range of analyses including: generalized additive models, linear and non-linear modelling, regression trees, Bayesian statistics etc. Currently there are over 2500 packages available and this number continues to grow. These packages are readily available through a global network of repositories called the Comprehensive R Archive Network (CRAN).

2.3. Openness and reproducibility

A key characteristic of our approach was to ensure the software was fully open to scrutiny. There were several reasons for this decision. First, it is likely in the longer term that *trustworthy* software will be developed if it is fully open to scrutiny. Second, an open system encourages other developers to contribute to the project. In the UK and Europe there is considerable reliance on proprietary software for environmental modelling and assessment. Many argue that proprietary models should be discouraged in environmental regulation, believing that those affected by decisions where environmental models and tools are used should be able to scrutinize them (NAS, 2007). The development of *openair* is very much in this spirit. The choice of R as the platform for development is also highly appropriate because R itself and all R packages are open-source.

Ensuring that all the examples in an R package run as expected is part of the system of package development and testing in R. Packages submitted to CRAN are required to go through extensive checks of both the code and documentation and to ensure all provided examples run as expected. These checks are made on a daily basis. This process ensures that all code runs exactly as intended. Package development also strongly encourages the use of comprehensive help files, thus formalising and structuring the different functions contained in a package.

Beyond the inherent checking processes there are other options to produce documentation that is entirely reproducible. We have adopted a system called *Sweave* for the comprehensive documentation that is available for *openair* (Leisch, 2002). Sweave is a form of literate programming (Knuth, 1983) that allows R code to be embedded in latex documents. These documents must be `run' to compile a final output — usually a pdf report containing all the outputs from an analysis such as plots, tables and statistics. The purpose is to create dynamic reports, which can be updated automatically if data or analyses change. The dynamic nature of the process is particularly useful for software that is under continual development. Rather than running code to produce an analysis e.g. a plot, which is then pasted into the documentation, the master document contains the R code necessary to produce it. When run in this way, all data analysis output (tables, graphs, etc.) are created during compilation and inserted into a final latex document automatically. The report can be easily updated if data or analysis change e.g. as modifications are made to the software, which allows for truly reproducible research. We have found that this process acts as another way of ensuring quality control of the code. It also ensures that users of the software can exactly reproduce all outputs as intended.

The *openair* project itself is developed using R-Forge (Theußl and Zeileis, 2009). R-Forge is a central platform for the development of R packages, R-related software and other projects. R projects are often developed on R-Forge before being published on CRAN. One of the key advantages for project development on R-Forge is the use of the version control system subversion, which offers a wide range of source code management tools (Pilato et al., 2004). Another benefit of R-Forge is that it allows other developers to contribute to code development, as is the case with *openair*.

The *openair* website at http://www.openair-project.org/ provides more information concerning the project and a comprehensive manual that supports the package.

3. Example of *openair* usage and capability

3.1. Introduction

This section demonstrates the usage of a few *openair* functions. A summary of most *openair* functions is shown in Table 1. While it is not possible to cover many functions in depth, the examples below highlight some of the underlying principles of usage. R itself works well in an interactive way e.g. an analysis is produced and the results from it suggest a refinement or point to the use of another function. Because the feedback to the user is almost immediate,